DOCUMENT RESUME

ED 148 893 TM 006 833

AUTHOR Stallings, Jane A.; Giesen, Philip A.

TITLE The Study of Reliability in Observational Data. Occasional Paper 19.

INSTITUTION Phi Delta Kappa, Bloomington, Ind. Center on Evaluation and Research.
PUB DATE Feb 77
NOTE 23p.; Revised version of a paper presented at the Annual Meeting of the American Educational Research Association (58th, Chicago, Illinois, April 15-19, 1974)
AVAILABLE FROM Phi Delta Kappa, Eighth and Union, Box 789, Bloomington, Indiana 47401 ($1.50)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Bias; *Classroom Observation Techniques; Classroom Research; *Codification; *Data Collection; *Error Patterns; Grade 1; Grade 3; Observation; Primary Education; Recordkeeping; *Reliability; *Research Problems; Scoring; Simulation; Video Tape Recordings
IDENTIFIERS *Observer Reliability

ABSTRACT

Observer reliability and the confusability of codes, two sources of error in the collection of classroom observational data, are examined. Confusability is defined as the extent to which one code is mistakenly recorded as another code. Observational data were collected in each of 172 first grade and 171 third grade Follow Through and comparison classrooms in urban and rural locations throughout the U.S. One section of the Stanford Research Institute's classroom observation instrument was analyzed for confusability of codes. Twenty simulated classroom situations were videotaped and coded by 63 trained observers. Matrices listing all the codes were constructed for each observer. The observer's coding was recorded so that their errors would be detectable, and coding errors were analyzed. Separate tables presented each observer's criterion video tape accuracy rate by sponsor, site, and grade level. Analyses of these matrices were also used to study the confusability of the codes. The results showed that the high rate of confusability of several observers could have been caused by overlapping code definitions, poor video tape examples, or inadequate training. (HV)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED148893

PHI DELTA KAPPA

EIGHTH AND UNION BOX 789
BLOOMINGTON, INDIANA 47401

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

PRESENTS AN

OCCASIONAL PAPER

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Phi Delta Kappa

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

TM006 833

CENTER ON EVALUATION DEVELOPMENT AND RESEARCH

CEDR

OCCASIONAL PAPER 19

THE STUDY OF RELIABILITY IN OBSERVATIONAL DATA

by

Jane A. Stallings
and
Phillip A. Giesen

Stanford Research Institute
Menlo Park, California

February 1977

$1.50

3

This particular paper is a revised and shortened version of a presentation originally given at a session of the American Educational Research Association 1974 Annual Meeting. The utility of the study is two-fold. First, the procedure used provides for differentiation between errors on the part of the observer and those resulting from ambiguity in behavior codes. Second, and on a broader scale, the methodology presented here is one that can lend greater precision and reliability to observational data of other types. Given the interest expressed in process evaluation of teaching, questions arise concerning the reliability of data from observational techniques. By focusing on the confusability of codes, this study attempts to resolve some of those questions.

William J. Gephart
Director, Center on Evaluation,
Development and Research
Phi Delta Kappa

A STUDY OF RELIABILITY IN OBSERVATIONAL DATA

Classroom observation as a measurement technique has the potential for obtaining valuable information regarding teacher and child behavior. This potential is realized to the extent that the observation data can be shown to be reliable. Any description of classrooms, of course, has many potential sources of error. Classroom processes most certainly will vary from day to day. The observers will vary somewhat in their accuracy in recording the data and the codes themselves may be overlapping and limited in describing what is actually happening. In the report of Stallings and Kaskowitz (1974) upon which this study is based, four sources of error were considered in the study of reliability: (a) observer accuracy in coding criterion videotapes, (b) the overlapping or confusable codes, (c) the uniformity of the observation data that was collected, and (d) the day-to-day variability of classroom process variables. This paper will examine only the first two sources of error, observer reliability and confusability of codes.

The Problem of Confusability

Confusability is defined as the extent to which one code is mistakenly recorded as another code. Most observation systems are based on a group of shorthand codes where a symbol, a letter, or a numeral will be operationally defined to stand for a person, an object, or an entire event. If that operational definition makes perfectly clear which events are to be included and which events are to be excluded in the code, no confusion will exist. Unfortunately for most coding systems, events in the real world are not always black or white; there are some grey areas. To the extent that coding definitions overlap and unusual events cannot easily be placed in one category or the other, there is the possibility of confusion. For example, in the SRI coding system referenced below, code "5," general comments, is defined as greetings, personal compliments, social or non-task related comments and remarks.

Child: "Hello, Ms. Brown."
Teacher: "Good morning, Jan. I like your hair."
Child: "My Mama curled it."

Teacher: "Your costume is really funny, Bill."
Child: "My Mama made it."

Code "6," task-related comments, is defined as a statement about the activity or task in which children or adults are involved.

(Situation: two children are working with clay, making animals for a zoo)

Child A: "This clay makes my hands sticky."
Child B: "The horn won't stay on my cow."
Child A: "Mine doesn't have a horn."

The previous two examples are easy to categorize as "5" (personal compliments) or "6" (task-related comments) according to the operational definitions. The problem comes when would-be general or social comments are related to a task. For example, Sally and Jan are role playing "beauty parlor." Sally has pinned and curled Jan's hair and has tied a ribbon in it. Now if Sally says "I like your hair, Jan," even though it is a personal compliment it could be coded a "6" because it is related to the task in which the children are involved.

The operational definitions as they stand do not suffice in this particular situation, and the problem of reliability cannot be attributed to inadequate observers but rather to the inadequate coding specifications. The problem can be solved to some extent by knowing which codes are not clearly defined and modifying the existing operational definitions so as to reduce the grey area where coding definitions do overlap.

Method of Approach

The current study was designed to collect observation data for three days in each of 172 first grade and 171 third grade Follow Through and comparison classrooms. The classrooms were examples of seven Follow Through models and were located in 36 projects which represented all regions of the country including both urban and rural locations. Two observers were hired for each project; one observed first grade and the other observed third

grade classrooms. Observations were conducted in the Spring of 1973 as a part of the national Follow Through program evaluation.

The instrument used was developed by Stanford Research Institute for use in the national evaluation of the Follow Through program. Only the interaction section of this observation instrument was analyzed for confusability of codes. The interaction instrument identifies the speaker, the person being spoken to, and the message being delivered. Each frame (see Figure 1) represents a sentence. If one person asks a question it is coded in the first frame. The response of the other person is recorded in the second frame. Seventy-two of these frames represent a five-minute interaction period.

All observers were trained in a seven-day training course. Only those who met the final criterion of recording all codes on a videotape at a reliability of 70% or above were allowed to observe for the national Follow Through evaluation. Sixty-three of the original 72 met this criterion.

Approximately two weeks after the observers had been working in the classrooms, 20 simulated classroom situations were videotaped and sent for the observers to code. Each simulation was approximately 20 interaction frames long. These simulations attempted to present several concise, clear examples of each code used in the SRI observation instrument. Each simulation began with a still frame where the narrator identified the focus of the observation (a teacher or a child). Each skit was shown with a two-second pause between each interaction. The observers were instructed to code one interaction frame during each stop or pause.

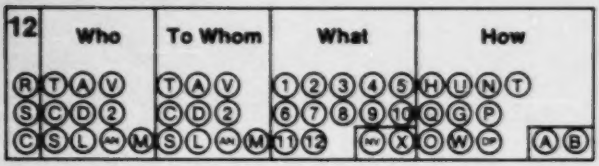
Matrices were constructed for each observer. A form was prepared which listed all 16 "What" codes across the top of the matrix and down the side. The reliability booklets for each observer were examined frame by frame. Tallies were made of each observer's coded interaction sequences. If the observer's coding agreed with the criterion, a tally was placed in the intersection of the row and column.* The principal diagonal then

*To illustrate the procedure, but for parsimony, only the data from the "What" code will be discussed in this paper.

FIGURE 1

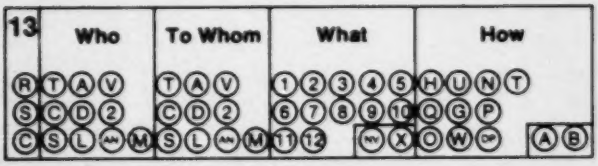
EXAMPLES OF INTER-ACTION FRAMES

1. Teacher: John, what is ten plus five plus five?



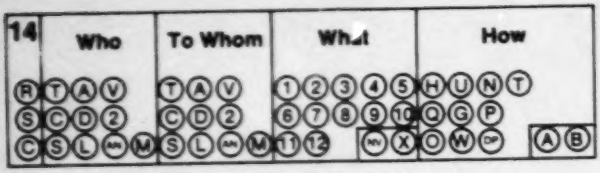
In our shorthand, this sentence is coded TCLQA. The teacher (T, Who Column) has asked a child (C, To Whom column) a direct question (1, What column; Q, How column) which is academic (A, How column).

2. John: Twenty.



Shorthand: CT3A. The child (C, Who column) responds (3, What column) with academic content (A, How column) to the teacher's (T, To Whom column) question.

3. Teacher: Very good, John.



Shorthand: TC8A. The Teacher (T, Who column) praises (8, What column) the child's (C, To Whom column) response (3, What column).

contains the cells indicating the observer's correct coding; other cells contain incorrect coding. The column totals are the total number of criterion examples shown on the videotape for each code; the row totals are the total number of times an observer recorded each code whether correctly or not. The column total should be compared with the diagonal to compute the observer's reliability for each code. An examination of a particular cell reveals whether the code was recorded correctly or incorrectly and, if recorded incorrectly, the row cells show exactly which codes were confused with one another.

Table 1
SRI "WHAT" CODES

1	–	Command or Request
1Q	–	Direct Question
2	–	Open-ended Question
3	–	Response
4	–	Instruction, Explanation
5	–	Comments, Greetings; General Action
6	–	Task-related Statement
7	–	Acknowledge
8	–	Praise
9	–	Corrective Feedback
10	–	No Response
11	–	Waiting
12	–	Observing, Listening
NV	–	Nonverbal
X	–	Movement

Observer Reliability

From Figure 2 we can see that this observer was quite accurate in her recording. Codes 1, 3NV, 4NV, 8 and 11 were perfect. She coded correctly 20 out of 22 "1Q's" 16 out of 17 "6s," and 22 out of 23 "12s." She coded "9s" and "1Q's" twice but her primary difficulty was in coding "5s" – social comments. She coded these as "6s" – task-related comments – five out of six times. It is possible that the observer is negligent or has failed to

FIGURE 2 – TOTAL NUMBER OF "WHAT" CODES RECORDED IN EACH CELL BY ONE OBSERVER

Code as Recorded by Observer (by Row)	*Criterion Examples (by Column)*																*Total No. Recorded by Observer*
"What" Codes	*1*	*1Q*	*2*	*3*	*3NV*	*4*	*4NV*	*5*	*5NV*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	
1	3																3
1Q		20														1	21
2		1	5														6
3		1		26							1						28
3NV					2												2
4						4											4
4NV							8										8
5								1		5							6
5NV									6								6
6			1							16							17
7											8						8
8											1	2					3
9		2											8				10
10														0			0
11															3		3
12																22	22
Total No. of Examples on Criterion Tape	3	22	6	27	2	7	8	6	7	17	11	2	8	0	3	23	

understand the code, or the training procedures were not adequate to help her see the fine discrimination between social comments and task-related comments.

Thus, observer bias can be assessed by examining the overuse, underuse, or confusion of codes. In this study, each observer was responsible for observing one grade level at a single site. Therefore, the data collected by each observer is identifiable in the analysis. The value of this method for measuring accuracy is that it contributes directly toward interpreting the data.

In order to determine the accuracy rates for each observer separately, tables were constructed to present each observer's criterion tape accuracy rate by sponsor, by site, and by grade level (see Table 2). Thus, for example, since the observer in Grade 1 at Location A had some difficulty with code "4," instruct, it is possible to examine those variables which use code "4" and compare the frequency of those variables reported at each of the five locations of the sponsor. If the frequency of the variable at the other four locations is similar and the frequency of Location A differs from the other four on those variables, there are two possible explanations: (1) Location A may be truly different from the other four sites, or (2) the observer of the first grades at Location A may not be recording accurately. In any case, the data resulting from code "4" at Location A would be interpreted with caution. The overall accuracy rate for observers was also computed and is presented in the righthand column of Table 2. This serves as a first screening of poor observing. Immediately it is evident that the observer in Location B third grades had lower ability than other observers. This procedure allows for each observer's data to be reviewed in order to estimate the accuracy of the individual on each code and to allow for the data to be interpreted accordingly.

A Study Comparing Paired Inter-Rater Reliability and Criterion Videotaped Reliability

Videotaped simulation of classroom events is, admittedly, different from actual classroom events. In an effort to compare the accuracy of observer ratings on the videotaped simulations and inter-rater accuracy of paired observations in the classrooms, a small study was conducted in one location. This section

TABLE 2 – ACCURACY RATES FOR THE "WHAT" CODES CODED BY OBSERVERS OF FAR WEST LABORATORY SITES

	Criterion Instances of "What" Codes														
	Six or More Examples									*Five or Less Examples*					
	1Q	*3*	*4*	*4NV*	*5NV*	*6*	*7*	*9*	*12*	*1*	*2*	*3NV*	*5*	*8*	*11*
Location A															
Grade 1	.84	.96	.62	.82	1.00	.93	.90	.86	.94	1.00	.57	.50	.00	.67	1.00
Grade 3	.81	1.00	1.00	1.00	.50	1.00	.90	.72	.82	.05	.60	.75	.50	.67	1.00
Location B															
Grade 1	.86	.96	.86	.83	.71	.62	.86	.71	.92	.75	.60	1.00	.09	.50	1.00
Grade 3	.67	.79	.67	.67	.80	.33	.60	.71	.94	.50	.43	.50	.20	.50	1.00
Location C															
Grade 1	.92	.81	.71	1.00	.74	.77	.88	.83	.91	1.00	.36	.80	.00	1.00	1.00
Grade 3	.94	.96	.78	.79	.86	.75	1.00	.88	.87	1.00	.50	1.00	.14	.67	1.00
Location D															
Grade 1	.80	.92	.83	1.00	.44	1.00	.90	.67	.75	.74	.40	.74	.10	.40	1.00
Grade 3	.78	1.00	1.00	.92	1.00	.74	.90	1.00	.79	1.00	.38	1.00	.13	.67	1.00
Location E															
Grade 1	.75	.94	.54	1.00	.89	.79	.86	1.00	.76	.60	.43	.00	.00	1.00	1.00
Grade 3	.95	.92	.75	.88	1.00	.58	.73	1.00	.93	1.00	.36	.00	.25	1.00	1.00

Accuracy Rate: Proportion correct of the total recorded.

compares the accuracy for two observers using both systems to evaluate observer reliability.

The first method, the paired observation, is the most commonly used method of assessing observer reliability. The procedure followed is to have the two observers situated in the same classroom, coding the same situation simultaneously. The recorded codes are then evaluated in terms of percent agreement between the two observers. If neither of the two observers has been designated as the criterion then it is difficult to know which set of data is correct or, worse yet, the two observers may be in agreement and both be unreliable. There are other serious limitations in paired observation. First, two extra people in the classroom are more obtrusive than one. Second, it is almost impossible to assure that the two observers are focusing on exactly the same action. Due to limited space, the two observers may not have the same angle of observation; thus, what they see and hear may be somewhat different and yet each observer could be collecting a correct and adequate sample of the behavior which is occurring. Lastly, it happens that during the paired observations certain interactions or codes may not occur, or occur at such a minimal rate, that reliability cannot be computed. There is no way to be certain that all codes will be assessed within a given time period.

In this study, an SRI staff trainer (designated as the criterion) observed with two field observers at their location. Data from 16 five-minute observations were examined, using 13 "What" codes. To assess the coding accuracy of the two observers, the proportion of frames that contained a particular code was recorded for the trainer and each trainee.

Table 3 shows the overall percentage reliability of the codes separately in terms of their ratio of frequency. It must be noted that accuracy for low frequency variables is difficult to interpret since if one observer records an event four times and the other only two times, and they observe an equal number of frames, the agreement is only 50%, even though the actual difference is only two occurrences. The data for both observers is presented on Table 2 in terms of percent agreement with the SRI trainer.

The results of each observer's coding of the criterion videotape of simulated classroom scenes are also presented on Table 3. The two sets of reliability data are extraordinarily similar.

TABLE 3 – PAIRED OBSERVATION RESULTS

	Observer 2		*Observer 1*	
"What" Code	*Percent Videotape Agreement*	*Percent Inter-Rater Agreement*	*Percent Videotape Agreement*	*Percent Inter-Rater Agreement*
1	50*	60	50*	88
1Q	73	73	86	90
2	100*	100**	100*	100**
3	82	89	89	88
4	80*	77	70*	99
5	100	92	–*	65
6	62	17**	80	81
7	60	75**	90	56**
8	–*	100**	100*	44**
9	67	55	100	65
10	–*	65**	–*	93**
11	100*	100**	100*	75**
12	91	70	100	83

*Fewer than 6 examples.

**Fewer than 60 criterion instances out of 1,152 opportunities (16 five-minute observations with 76 frames per observation).

Low occurring variables have a lower percent agreement on either procedure. For Observer 1 where there are sufficient examples of the codes, the two procedures are in close agreement about the observer reliability, i.e., 1Q, 3, 4, 6, and 12. Only code "9," corrective feedback, varies. On the videotape, Observer 1 coded all of the examples correctly while in the actual classroom situation she coded 60% of the examples according to the criterion. Observer 2 was low in her reliability in recording code "9" on both procedures.

While the videotaped events are limited in scope and differ from the classroom situation, they do offer a standard stimulus for examining each observer's ability to code specified events and to identify observer bias. There are some errors in the system;* however, the variation introduced by a second observer is eliminated. The two systems of examining observer accuracy do yield somewhat different information, yet the results are not contradictory. The videotape system is easier and less expensive to administer since by sending criterion tapes to observers, the cost of traveling a criterion person from one location to another is eliminated.

Confusion of Observation Codes
(A Study Combining the Results of
63 Observers)

Analyses of these matrices were also used to study the confusability of one code with another. By combining the results for all observers, the extent of general confusability of each code could be examined. Codes that reveal a high rate of confusion by several observers suggest these possible causes: an overlapping of code definitions, poor videotape examples, or less-than-adequate training procedures.

How the 63 observers coded the videotaped examples is shown in Figure 3. The diagonal shows the number of correct codings. The row at the bottom of the table is the number of videotape criterion examples coded by the combined observers. Each figure

*System errors include: faulty tapes, machines, examples of codes.

FIGURE 3 – TOTAL OF "WHAT" CODES RECORDED IN EACH CELL BY 63 OBSERVERS

Codes as Recorded by Observers "What" Codes	*Criterion Examples* 1	1Q	2	3	3NV	4	4NV	5	5NV	6	7	8	9	10	11	12	*Total No. Recorded by Observer*
1	160	13	2	5	1	3	0	3	0	9	1	0	13	0	0	12	222
1Q	27	966	135	24	1	10	0	0	0	4	4	0	29	0	0	12	1212
2	0	153	156	3	1	0	0	0	0	0	1	0	3	0	0	0	317
3	1	35	4	1310	17	2	0	4	2	29	14	0	10	0	0	19	1447
3NV	2	0	0	14	142	2	0	0	9	4	4	0	0	0	0	3	180
4	2	9	0	34	1	344	11	0	1	7	8	0	8	0	0	55	480
4NV	2	1	0	5	2	3	480	0	31	2	13	2	0	0	2	5	548
5	14	7	0	18	1	2	19	37	39	200	3	3	16	0	0	6	365
5NV	1	1	0	0	3	2	96	2	343	6	5	2	0	0	17	1	479
6	27	31	22	75	1	21	11	5	3	581	36	0	29	0	0	11	853
7	0	3	1	38	1	7	0	0	0	1	455	5	10	0	0	6	527
8	0	0	0	7	0	1	0	0	0	4	22	106	2	0	0	3	145
9	3	3	0	9	0	13	1	2	0	23	2	1	378	0	0	4	439
10	1	3	0	1	11	1	0	0	6	0	0	0	0	0	3	5	31
11	0	0	0	1	0	0	0	0	5	0	0	0	0	0	125	1	132
12	5	27	0	28	6	62	4	3	19	11	37	2	4	0	29	1190	1427
Number of Examples Times All Observers	245	1252	320	1572	188	473	622	56	458	881	605	121	502	0	176	1333	8804

in the bottom row can be compared with the corresponding cell in the diagonal. For example, code "1" was recorded correctly 160 times out of a possible 245 times. The other entries in the "1" row are sources of confusion.

The proportion of times that the observers were correct in their recordings is presented in Figure 4. For example, the observers recorded "1Q" correctly 80% of the time. Looking across the "1Q" row, we see that 2% of the time when a "1Q" was recorded it was truly a "1," and 11% of the time it was truly a "2." While there were only five examples of the "2" code, the 63 observers mistakenly coded the examples of "2" as a "1" 48% of the time. This finding suggests that an overlap in definitions exists between the "1" and the "2" code or that the videotape examples of the "2" code were not clear or were poorly executed.*

From this table we see a situation similar to that of the single observer discussed previously. The "5" code – social comments – was coded 55% of the time as a "6" code – task-related comments. This indicates that the problem lies in the definition of the codes or the examples of the codes rather than with the individual observer. All other codes are reliable at 70% and above for the combined 63 observers.

The confusability study identifies the specific codes that appear to be reliable as well as those that are confused and need to be redefined. The findings suggest that some codes, such as "5" and "6," should be more carefully defined. At present their definitions are overlapping. There were some difficulties in coding "2s;" however, there are not enough examples of code "2" on the tape to draw conclusions about how reliable the definition is since the source of error may be in the taped examples themselves.

Utility of the Method

Through this procedure instrument developers and trainers of observers receive feedback by examining the extent to which one

*A careful examination of the videotapes revealed that the quality of the tape was poor and the examples were not clear on two of the five skits exemplifying the "2" code.

FIGURE 4 – PROPORTIONAL DISTRIBUTION OF CODES RECORDED BY 63 OBSERVERS

Codes as Recorded by All Observers	*Criterion Examples*															
"What" Codes	*1*	*1Q*	*2*	*3*	*3NV*	*4*	*4NV*	*5*	*5NV*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
1	.72*	.06											.06			
1Q	.02	.80	.11													
2		.48	.48													
3		.02		.91						.02						
3NV				.08	.79											
4				.07		.72										.11
4NV							.88		.06							
5				.05			.05	.10	.11	.55			.04			
5NV							.20		.72							
6	.03		.03	.09						.68	.04		.03			
7				.07							.86					
8											.15	.73				
9										.05			.86			
10														.00		
11															.95	
12																.83
Total Number of Examples on Criterion Tape	4	20	5	26	3	8	11	1	7	16	11	2	9	0	3	23

*Only numerals in the diagonal report proportion correctly coded.

code is confused with another code. Operational definitions and training procedures can be made more precise. Also, the data reported by one observer can be systematically assessed for reliability. All of an observer's data need not be discarded if only one or two codes are less than acceptably reliable.

This procedure for assessing observer reliability and the confusability of codes could be used with other observation systems. As long as a criterion has been established a matrix can be formed of what ought to be coded and what is actually coded. This procedure provides more useful information than does the method most often used to report observer reliability, that is, comparison of two observers' marginal frequency counts of codes. Marginal frequency counts report the percent of time a code occurs out of all possible codes and the percent of two observers recording simultaneously is compared. The sequence of codes or the error source is not assessed. Therefore it is difficult to know whether the problem is in the coding definition or the observers.

Too often observational data has been considered a soft and unreliable measurement tool. In our opinion it is necessary and possible to raise the level of observational data so that it becomes a precision tool in evaluating instructional processes and specific elements in child growth and development (e.g., autonomy and social-emotional growth). This study in confusability hopefully is one step toward the goal of improving the precision of observational measurement.

OCCASIONAL PAPERS

1. THE PROBLEM AND PROBLEM DELINEATION TECHNIQUES – William J. Gephart, Phi Delta Kappa. Presented at the Second National Symposium for Professors of Educational Research, sponsored by Phi Delta Kappa, Boulder, Colorado, November 21, 1968. A discussion of the nature of the concept "problem" as related to educational research with a discussion of several techniques useful in problem identification and delineation. $1.00
2. A REVIEW OF INSTRUMENTS DEVELOPED TO BE USED IN THE EVALUATION OF THE ADEQUACY OF REPORTED RESEARCH – Bruce B. Bartos, Phi Delta Kappa & Indiana University. Presented at the Annual Meeting of the American Educational Research Association, February, 1969, Los Angeles, California. A brief description and bibliographic annotation of 40 instruments developed to be used in assessing the methodological quality of completed research. $.25
3. PROFILING EDUCATIONAL RESEARCH – William J. Gephart, Phi Delta Kappa, January, 1969. The rationale for the development of a methodology profile on completed research to show its strengths and weaknesses. Included are flow charts for profiling the five facets of the research process. $.75
4. APPLICATION OF THE CONVERGENCE TECHNIQUE TO READING – William J. Gephart, Phi Delta Kappa, January, 1969. An interim report on a research program planning effort in the field of reading. (Out of Print)
5. THE CONVERGENCE TECHNIQUE AND READING: A PROGRESS REPORT – William J. Gephart, Phi Delta Kappa. Presented at the Annual Meeting of the International Reading Association, May 2, 1969, Kansas City, Missouri. A second interim report on the planning of a reading research program. (Out of Print)
6. THE EIGHT GENERAL RESEARCH METHODOLOGIES: A FACET ANALYSIS OF THE RESEARCH PROCESS – William J. Gephart, Phi Delta Kappa, July 14, 1969. The identification and description of general research methods in education through the use of Gutman's facet design and analysis technique. It also details the procedures for the Gutman technique. This paper was printed in the proceedings of the Warsaw, Poland Congress of the International Association for the Advancement of Educational Research. $1.00
7. PROFILING INSTRUCTIONAL PACKAGE – William J. Gephart & Bruce B. Bartos, Phi Delta Kappa, August, 1969. An instruction text to assist individuals with no prior research training in the use of research profiling flow charts to assess the methodological adequacy of completed research. $1.00
8. EDUCATIONAL KNOWLEDGE USE – Gene V. Glass, Laboratory of Educational Research, University of Colorado. An analysis of the availability and use of empirically based information in education. $.50
9. MEASUREMENT AND RESEARCH IN THE SERVICE OF EDUCATION – Warren G. Findley, Research and Development Center in Educational Stimulation, University of Georgia. Originally presented as an invited address at the annual meeting of the American Educational Research Association; this paper uses an historical perspective to examine the role of measurement and research in education. $.75

10. THE EDUCATIONAL CATALYST: AN IMPERATIVE FOR TODAY – Joe H. Ward, Jr., Reeve Love, and George M. Higginson, Southwest Educational Development Laboratory, Austin, Texas, July, 1971. An analysis of the problems involved in the process of change and improvement of the practice of education. This paper poses a new professional speciality for the facilitation of empirically based educational improvements. $1.00
11. DISSERTATIONS YOU MAY WANT TO SEE – William J. Gephart, Phi Delta Kappa, 1970. A collection of dissertations done in 1969 which focus on research training. $.25
12. THE DOCTORATE IN EDUCATION IN CANADA – Neville L. Robertson, Commission on Higher Education, Phi Delta Kappa, 1971. An analysis of the institutions offering the doctorate in education in Canada. This paper is a companion piece to a larger study of similar institutions in the United States. $.75
13. THE IMPORTANCE OF STATISTICAL POWER IN EDUCATIONAL RESEARCH – John K. Miller, and Thomas R. Knapp, University of Rochester. When an educational experiment results in non-significant differences can it be said that no difference exists? This paper discusses the concept that must be attended to IF that question is to be answered. It also details the procedure for determination of sample size needed in an experiment. $1.25
14. THE CONDUCT OF DEVELOPMENT IN EDUCATION – Richard E. Schutz, SWRL Educational Research and Development. The past decade has developed awareness of the tools and procedures needed to perform educational tasks. Some have argued that the research process is the vehicle for achieving that end. Schutz argues that the development process is different from the research process in many important ways that need to be understood by educators. $1.50
15. THE PROCESS OF DEVELOPMENT – Gerald Nadler and William J. Gephart. The development process is examined from five points of view in this interview of Gerald Nadler, Chairman of the Industrial Engineering Department, University of Wisconsin. Questions posed by Gephart and their discussion examine the general nature of the process, the types of problems for which it is an appropriate resolution strategy, the component activities of the development process, the types of products resulting from it, and the criteria that would be used in assessing the quality of a development effort. $1.50
16. THE MOST SIGNIFICANT EDUCATIONAL RESEARCH CONTRIBUTIONS OF THE PAST TEN YEARS – William J. Gephart, Phi Delta Kappa, in collaboration with Berlie Fallon, Anton J. Netusil, and William K. Poston, Jr. Three individuals were asked to identify important research contributions. Work thus identified is briefly summarized and comments are made about the manner in which educational research *ought* to be looked to by educators. $1.25
17. EVALUATION: PAST, PRESENT, AND FUTURE – William J. Gephart, Phi Delta Kappa. Evaluation's history is traced from the movements in measurement, in the Eight-Year Study, and in accreditation. Next evaluation is examined as it has grown from 1965 to 1975. A schema is presented for integrating the many, many models of evaluation that have come into being during that decade. Finally, the paper suggests some possible directions for evaluation's future given current focus in the field. $1.50.

18. THE EVALUATION OF TEACHING – NSPER: 74 – William J. Gephart, Robert B. Ingle, and Gary Saretsky, Phi Delta Kappa, 1976, 104 pages. Alternative methods for evaluating teaching, their strengths and weaknesses, and the nature of teaching are the focii of this report. It covers the sessions of the National Symposium of Professors of Educational Research held in Denver, Seattle, and Atlanta. Central to these sessions was a one-day workshop called TAFI (Teacher Appraisal for Improvement), which acquaints participants with eight methodological approaches for evaluating teaching. $2.00